Learning partial ordinal class memberships in a proportional odds setting

Jan Verwaeren JAN.VERWAEREN@UGENT.BE
Willem Waegeman WILLEM.WAEGEMAN@UGENT.BE
Bernard De Baets BERNARD.DEBAETS@UGENT.BE
KERMIT, Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, Ghent, Belgium

Keywords: proportional odds models, partial class membership, kernel methods, ordinal regression

Abstract

In a typical classification problem, each data object is restricted to belong to a single class. However, in settings where the classes are defined in a non-crisp way, this might be too conservative. As a solution, we present an approach which allows data objects to exhibit a degree of membership to several classes. More specifically, we shall consider the case where the set of classes is equipped with a linear order. We describe statistical models which could underlie this kind of data as well as algorithms that can be used to learn input-output relations in such a non-crisp case.

1. Introduction

In the traditional binary or multi-class classification setting, usually the restriction is made that data objects belong to a single class, i.e., to every data object one can associate a single label from a finite unordered set of class labels. In numerous applications, however, this setting can be considered too conservative. Consider for instance the well-studied problem setting of multi-label classification, in which not a single label but a set of labels is associated to every data object. Similarly, partial membership models can also be seen as a generalization of traditional multi-class classification, in which to every data object a membership degree to every class is associated, instead of a crisp class label.

Data consisting of partial class memberships can be found in many domains, such as text categorization, social network analysis, microbiology and agriculture. As a result, researchers in statistics, machine learning and fuzzy set theory have shown interest in developing learning algorithms for partial class memberships. However, previous research mainly focused on unsupervised learning algorithms, like clustering algorithms such as the fuzzy c-means algorithm, where a data instance can simultaneously exhibit a degree of membership to several clusters.

Recently, a supervised machine learning algorithm was proposed for multi-class classification problems where partial class memberships are observed (Anonymous, 2009). In this paper, a similar setting is considered but in addition it is assumed that a linear order is specified on the classes, which naturally follows from the semantics of these classes (e.g. bad, moderate, good). This leads to an ordinal regression setting where the labels consist of partial class memberships. Compared to multi-class classification, the presence of a linear order results in two important challenges for developing partial class membership models. These two challenges directly follow from the two main differences between multi-class classification and ordinal regression:

1. Firstly, the absence or presence of a linear order on the classes gives rise to a different model structure for the two types of problems. In contrast to their multi-class counterparts, models for ordinal responses typically assume and approximate an underlying latent variable that reflects the order on the classes.

2. Secondly, multi-class classification and ordinal regression models typically differ in the type of performance measure they optimize. If a linear order

Appearing in *Proceedings of the 19th Machine Learning conference of Belgium and The Netherlands.* Copyright 2010 by the author(s)/owner(s).

on the classes can be assumed, then a performance measure that takes this order into account must be utilized, both for optimization and evaluation.

2. Ordinal regression

Since our methodology can be seen as an extension of ordinal regression, we give a small review of some key principles of it. Let us use the familiar multi-class classification setting (ordinal regression can be seen as a special case) to introduce some notations. The goal is to learn a mapping from an input space $\mathbb{X}$ to a finite set $\mathcal{C} = \{C_1, \ldots, C_K\}$ containing K labels. To this end, each object is usually represented by a D-dimensional feature vector $\mathbf{x} \in \mathbb{X}$ and a class label $y \in \mathcal{C}$. A training dataset $\mathbf{T}$ of N i.i.d. observations can then be denoted as a set of couples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ with $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,D})$, in which we assume that the couples $(\mathbf{x}_i, y_i)$ are realizations of the random vector $(\mathbf{x}, y)$. Moreover, using this notation, an ordinal regression problem can be seen as a special case of a classification problem where the label set is endowed with a linear order $C_1 \prec C_2 \prec \ldots \prec C_K$.

In a standard machine learning setting, one aims to find a mapping or model $f : \mathbb{X} \to \mathcal{C}$ that minimizes the expected value of some regularized loss function, i.e.

$$\hat{f}(\mathbf{x}) = \min_{f \in \mathcal{H}} \mathcal{L}(f, \mathbf{T}) + \lambda J(f)\,, \tag{1}$$

with $\mathcal{L}$ a loss function on the training dataset, $\mathcal{H}$ a hypothesis space of models, J a penalty term for the complexity of the model and λ a regularization parameter.

As mentioned in the introduction, the presence of an order relation on the class labels has implications on the model structure and on the type of loss function that is used. Both of these implications are briefly described next.

2.1. The latent variable motivation

Roughly speaking, an ordinal regression model $f : \mathbb{X} \to \mathcal{C}$ maps a data object to one of the classes of $\mathcal{C}$. The vast majority of existing ordinal regression models can be represented in the following general form:

$$f(\mathbf{x}) = \begin{cases} C_1 & \text{, if } g(\mathbf{x}) \leq \theta_1 \\ C_2 & \text{, if } \theta_1 < g(\mathbf{x}) \leq \theta_2 \\ \vdots & \\ C_K & \text{, if } \theta_{K-1} < g(\mathbf{x}) \end{cases} \tag{2}$$

with $\theta_1, ..., \theta_{K-1}$ free parameters and $g : \mathbb{X} \to \mathbb{R}$ any function that assigns a real value to a data object. The function g allows to impose an ordering on a collection of data objects and it is therefore in machine learning often referred to as a *ranking* function or latent variable function. The use of a latent variable can be motivated by interpreting an ordinal scale as the result of course measurements on a continuous scale. As a consequence, the ordinal classes correspond to non-overlapping intervals covering the entire real line. Because of this, fitting an ordinal regression model can be subdivided into two parts: the functional form of the latent variable has to be estimated on the one hand and the thresholds that define the non-overlapping intervals have to be chosen on the other hand.

The proportional odds model (McCullagh, 1980) without doubt the best known and most applied technique to represent ordinal responses, it follows naturally from the latent variable interpretation. Instead of fitting a decision rule $f : \mathbb{X} \to \mathcal{C}$, this type of model defines a probability density function over the class labels for a given feature vector $\mathbf{x}$. The cumulative probability $\overline{p}_k$ of observing a label smaller than or equal to C_k is defined as follows:

$$\overline{p}_k(\mathbf{x}) = \mathcal{P}\{y \leq C_k \mid \mathbf{x}\}\,,$$

with $(\mathbf{x}, y)$ an instance-label couple. In general, it is assumed that

$$\mathcal{P}\{y \leq C_k \mid \mathbf{x}\} = \phi(g(\mathbf{x}) + \theta_k)\,,$$

for $k = 1, ..., K-1$; with $\phi(\cdot)$ any cumulative distribution function and θ_k the threshold parameter for class C_k. When trying to fit this model to a dataset, it is (in its most basic form) assumed that $g(\mathbf{x})$ can be written as a linear combination $\mathbf{w} \cdot \mathbf{x}$ of the inputs, where $\mathbf{w}$ is a coefficient vector.

2.2. Performance measures

Another important difference between multi-class classification and ordinal regression can be found in the loss function used to optimize and evaluate the model. To evaluate the performance of a given multi-class classification model, the accuracy on a test dataset is typically measured, and a differentiable approximation of accuracy such as the logistic loss or hinge loss is typically optimized on training data to fit the parameters of the model. In an ordinal setting however, the use of accuracy seems unnatural. For instance, when the class labels are {*bad*, *moderate*, *good*}, classifying a *good* instance as *bad* is worse than classifying it as *moderate*. As a result, other performance measures such as the concordance index have been proposed in literature.

3. Partial class memberships

In the previous section, objects were restricted to belong to a single class. In this section this restriction is relaxed, resulting in partial class memberships. To this end, each object will be linked with a K-dimensional real-valued vector that will be called its partial membership vector. Each object has one unit of membership, divided over the K classes. As a result, each component of the partial membership vector is positive and the sum of all components of the vector equals one (a type of data often referred to as compositional data in the statistical literature (Aitchison, 1986)). In a K-class problem, a partial membership vector $\mathbf{y}$ is a vector within the K-dimensional simplex $\mathcal{Y}^K$:

$$\mathcal{Y}^K = \Bigg\{ \mathbf{y} = (y_1, \ldots, y_K) \in \mathbb{R}^K \mid y_i \geq 0, \\ \forall\, i \in \{1, \ldots, K\}; \sum_{i=1}^{K} y_i = 1 \Bigg\} . \quad (3)$$

The i-th instance in a training set $\mathbf{T}$ will now be denoted

$$(\mathbf{x}_i, \mathbf{y}_i) = \Big((x_{i,1}, \ldots, x_{i,D}), (y_{i,1}, \ldots, y_{i,K}) \Big) .$$

As another extension of the crisp setting, the predictive model that we aim to fit to the data will be represented as $\mathbf{f} : \mathbb{X} \rightarrow \mathcal{Y}^K$, in which $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_K(\mathbf{x}))$, where we require $\sum_{k=1}^{K} f_k(\mathbf{x}) = 1$. We could then choose $\mathbf{f}$ as a set of parameterized functions for which the parameters are estimated through optimization of the loss function

$$L(\mathbf{f}(\mathbf{x}), \mathbf{y}) = \sum_{k=1}^{K} (f_k(\mathbf{x}) - y_k)^2 .$$

The choice of such a model seems justified since, when there exists no order on the class labels, the following multivariate model $\mathbf{f}$ can be assumed to underly the data

$$\mathbf{y} = \begin{cases} y_1 &= f_1(\mathbf{x}) + \gamma_{1,\mathbf{x}} \\ &\vdots \\ y_K &= f_K(\mathbf{x}) + \gamma_{K,\mathbf{x}} \end{cases} \quad (4)$$

where $\gamma_{1,\mathbf{x}}, \ldots, \gamma_{K,\mathbf{x}}$ are (dependent) error terms with zero mean (the random part of the model) and the following properties hold:

$$\sum_{k=1}^{K} y_k = \sum_{k=1}^{K} f_k(\mathbf{x}) = 1 \quad , \forall \mathbf{x} \in \mathbb{X} , \quad (5)$$

$$f_k(\mathbf{x}) + \gamma_{k,\mathbf{x}} \geq 0 \quad , k = 1, ..., K; \ \forall \mathbf{x} \in \mathbb{X} , \quad (6)$$

$$\sum_{k=1}^{K} \gamma_{k,\mathbf{x}} = 0 \quad , \forall \mathbf{x} \in \mathbb{X} , \quad (7)$$

$$\mathbf{E}[\mathbf{y} \mid \mathbf{x}] = \mathbf{f}(\mathbf{x}) \quad , \forall \mathbf{x} \in \mathbb{X} , \quad (8)$$

$$\mathbf{E}_{\mathbf{x}}[\gamma_{k,\mathbf{x}}] = 0 \quad , k = 1, ..., K . \quad (9)$$

Given a full description of the components of (4) is not a trivial task. First, let us reformulate this problem. Suppose we observe an object with feature vector $\mathbf{x}$, whose partial membership vector is modeled by (4). Which probability density functions would then be possible for $\mathbf{y}$? To be able to answer this question, a set of real-valued functions $\mathbf{f}$ respecting (5) is needed, as well as a complete description of the dependence structure of $\gamma_{1,\mathbf{x}}, \ldots, \gamma_{K,\mathbf{x}}$. Because of constraints (6)–(9), the latter is not a simple task. As such, it is not our aim to determine all possible distributions of $\mathbf{y}$. Instead, we present an example of a setting in which the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$, further denoted as $\mathbf{y} \mid \mathbf{x}$, respects (5)–(9).

To start, consider a fixed feature vector $\mathbf{x}$. Plugging this feature vector into model (4) should produce a random vector with a fully described probability density function $\mathbf{y} \mid \mathbf{x}$ over the simplex satisfying (5)-(9). A set of suitable functions $f_1, \ldots, f_K$ can easily be found. Recall constraint (8), which requires that $\mathbf{E}[\mathbf{y} \mid \mathbf{x}] = (f_1(\mathbf{x}), \ldots, f_K(\mathbf{x}))$. To accomplish this, we can for instance use the Dirichlet distribution with its parameter set $\boldsymbol{\beta}$ conditioned on $\mathbf{x}$. The parameter set $\boldsymbol{\beta}(\mathbf{x}) = (\beta_1(\mathbf{x}), \ldots, \beta_K(\mathbf{x}))$ associated with the feature vector $\mathbf{x}$ is obtained as

$$\beta_k(\mathbf{x}) = s\, f_k(\mathbf{x}) \quad , \text{for} \quad k = 1, \ldots, K ,$$

with s a positive real parameter. It can easily be shown that such a model will respect constraint (9). In general, using the Dirichlet distribution, the probability density function of the partial class membership vector conditioned on the inputs takes the following form:

$$\mathbf{y} \mid \mathbf{x} \sim \mathrm{Dir}\Big(\beta_1(\mathbf{x}), ..., \beta_K(\mathbf{x}) \Big) . \quad (10)$$

4. Partial ordinal class memberships

Model (4) can be used to describe partial ordinal class memberships as well. However, when doing so, the ordinal nature of the class labels is lost. In this section

we will show that, through the addition of some extra constraints on (4), the information provided by the order on the classes can be preserved and incorporated into the model.

4.1. Ordinality w.r.t. partial memberships

When the classes are ordered, this order is incorporated into the model through the use of a latent variable. When considering the proportional odds model, it can be seen that the highest classes (in the linear order) become more likely as the value for the latent variable increases. This property is key to any ordinal regression model. As a consequence, we want to have a similar property in our ordinal partial membership model. However, the general problem setting does not define a linear order on the partial membership vectors as it does on the classes. As a result, we will have to define/assume a type of order on the set of partial membership vectors which reflects the order on the class labels. In this paper, we will assume a type of order relation which is strongly related to the concept of first order stochastic dominance (Levy, 2006). First, let us introduce the notion of a cumulative partial membership vector.

Definition 1 *For the K-dimensional partial membership vectors* $\mathbf{y}$ *and* $\mathbf{f}(\mathbf{x})$*, the cumulative partial membership vectors* $\overline{\mathbf{y}} = (\overline{y}_1, \ldots, \overline{y}_K)$ *and* $\overline{\mathbf{f}}(\mathbf{x}) = (\overline{f}_1(\mathbf{x}), \ldots, \overline{f}_K(\mathbf{x}))$ *are defined as*

$$\overline{y}_k = \sum_{\ell=1}^{k} y_\ell \quad \textit{and} \quad \overline{f}_k(\mathbf{x}) = \sum_{\ell=1}^{k} f_\ell(\mathbf{x}) \,,$$

for $k = 1, \ldots, K$*.*

The notion of cumulative partial membership vectors can then be used in the following definition.

Definition 2 *Given two partial membership vectors* $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}^K$*, we say that* $\mathbf{y}_1$ *dominates* $\mathbf{y}_2$ *(denoted* $\mathbf{y}_1 \succcurlyeq_{SD} \mathbf{y}_2$*) if the following holds for the cumulative partial membership vectors* $\overline{\mathbf{y}}_1 = (\overline{y}_{1,1}, \ldots, \overline{y}_{1,K})$ *and* $\overline{\mathbf{y}}_2 = (\overline{y}_{2,1}, \ldots, \overline{y}_{2,K})$*:*

$$\forall k \in \{1, \ldots, K\} : \overline{y}_{1,k} \leq \overline{y}_{2,k} \,.$$

We say that $\mathbf{y}_1$ *strictly dominates* $\mathbf{y}_2$ *if*

$$\mathbf{y}_1 \succcurlyeq_{SD} \mathbf{y}_2 \quad \textit{and} \quad \mathbf{y}_1 \neq \mathbf{y}_2 \,. \tag{11}$$

It can easily be seen that the dominance principle defines a partial order relation on a set of K-dimensional partial membership vectors. We will use this partial order relation, combined with the latent variable interpretation, to construct a model that can underly partial class membership data on an ordinal scale.

Our model can be seen as a two-step process. Firstly, the feature vector $\mathbf{x}$ of an object is mapped to a real number, being the object's value for the latent variable. Secondly, a mapping is performed from this latent variable to the space of partial membership vectors. As before, a latent variable $g(\mathbf{x})$ can be used to construct a model $\mathbf{f} : \mathbb{X} \to \mathcal{Y}^K$ as follows

$$\mathbf{f}(\mathbf{x}) = \mathbf{h}(g(\mathbf{x})) \,, \tag{12}$$

where $\mathbf{h} = (h_1, ..., h_K) : \mathbb{R} \to \mathcal{Y}^K$ will be called link functions. Through the addition of random components ϵ and $\gamma_{1,\mathbf{x}}, \ldots, \gamma_{K,\mathbf{x}}$, an ordinal variant of (4) is obtained:

$$\mathbf{y} = \begin{cases} y_1 &= h_1(g(\mathbf{x}) + \epsilon) + \gamma_{1,\mathbf{x}} \\ &\vdots \\ y_K &= h_K(g(\mathbf{x}) + \epsilon) + \gamma_{K,\mathbf{x}} \end{cases} \tag{13}$$

satisfying (5)–(9).

In an ordinal regression setting, the latent variable motivation suggests that a monotone relationship exists between the latent variable and the output (the predicted class label). In our partial ordinal class membership model, we want to preserve this monotone relationship between the latent variable and the output (which is in this case a partial membership vector). To be able to speak of a monotone relationship between a latent variable and a partial membership vector, we will use the order implied by the dominance principle described before. This results in the additional constraint given in the following definition.

Definition 3 *The model* $\mathbf{f}$ *defined in (12) is called monotone with respect to the dominance principle if for any two feature vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *the following equivalence holds:*

$$\mathbf{f}(\mathbf{x}_1) \succcurlyeq_{SD} \mathbf{f}(\mathbf{x}_2) \Leftrightarrow g(\mathbf{x}_1) \geq g(\mathbf{x}_2) \,.$$

Requiring a model to be monotone with respect to the dominance principle imposes some constraints on $\mathbf{h}$. Different sets of constraints might lead to monotone models and it is not our intention to present all possibilities here. Instead, a simple method is presented which will ensure monotonicity. Let us start by defining the cumulative counterpart of $\mathbf{h}$, as $\overline{\mathbf{h}} = (\overline{h}_1, \ldots, \overline{h}_K)$ where

$$\overline{h}_k(u) = \sum_{l=1}^{k} h_l(u) \quad \text{, for} \quad k = 1, \ldots, K; \quad \forall u \in \mathbb{R},$$

similar to what we did before for $\mathbf{y}$ and $\mathbf{f}$. It can easily be seen that the vector of functions $\overline{\mathbf{h}}$ uniquely defines the vector of functions $\mathbf{h}$ and vice verse. As a result, a set of constraints imposed on $\mathbf{h}$ can be translated into a set of constraints on $\overline{\mathbf{h}}$ and vice verse. This duality is exploited in the following obvious proposition which states that requiring $\overline{h}_1, \ldots, \overline{h}_{K-1}$ to be decreasing functions suffices to obtain a monotone model.

Proposition 4.1 *The model* $\mathbf{f}$ *defined in (12) is monotone with respect to the dominance principle if* $\overline{h}_1, \ldots, \overline{h}_{K-1}$ *are decreasing functions.*

As for model (4), we can look at (13) from a distributional point of view and use it to define a conditional distribution function $\mathbf{y} \mid \mathbf{x}$. In this context we argue that being monotone with respect to the dominance principle is desirable for any model used to described partial ordinal class memberships. Consequently, it is desirable to retain this property when designing conditional probability density functions for $\mathbf{y} \mid \mathbf{x}$. This is formalized in the following definition.

Definition 4 *Model (13) is called a monotone distribution generating model with respect to the dominance principle if for any two feature vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*, the following equivalence holds*

$$\mathbf{E}[\mathbf{y}_1 \mid \mathbf{x}_1] \succcurlyeq_{SD} \mathbf{E}[\mathbf{y}_2 \mid \mathbf{x}_2] \Leftrightarrow g(\mathbf{x}_1) \geq g(\mathbf{x}_2)\,,$$

where $\mathbf{y}_i \mid \mathbf{x}_i$ *represents the output of model (13) conditioned on an input* $\mathbf{x}_i$*.*

The following proposition presents a setting in which (13) is a monotone distribution generating model with respect to the dominance principle.

Proposition 4.2 *Model (13) is a monotone distribution generating model with respect to the dominance principle if the following properties hold:*

(i) $\overline{h}_1, \ldots, \overline{h}_{K-1}$ *are decreasing functions,*

(ii) $\mathbf{E}[\epsilon] = 0$*,*

(iii) The conditional distribution of $\mathbf{y} \mid \mathbf{x}$ *is a Dirichlet distribution with parameter set* $\boldsymbol{\beta}(\mathbf{x}) = (\beta_1(\mathbf{x}), \ldots, \beta_K(\mathbf{x}))$ *where*

$$\beta_k(\mathbf{x}) = s\, h_k(g(\mathbf{x}) + \epsilon), \quad \text{for} \quad k = 1, \ldots, K; \quad \forall \mathbf{x} \in \mathbb{X}$$

where s *is a positive real parameter.*

4.2. Performance measures

In the previous section, a general structure for ordinal partial class membership models was developed. If the performance of a model $\mathbf{f}$ has to be evaluated, a performance measure respecting the ordinal nature of the data is needed. To this end, we choose the mean absolute error on the cumulative partial membership vectors as a performance measure.

$$L_{\text{MAEC}}(\mathbf{f}(\mathbf{x}), \mathbf{y}) = \sum_{k=1}^{K} \left| \overline{f}_k(\mathbf{x}) - \overline{y}_k \right| .$$

4.3. Learning ordinal partial class memberships

In a machine learning context, the aim is to learn a model of type (12), satisfying the constraints given in Proposition 4.1. The proportional odds model as introduced in Section 2.1 naturally establishes a monotone relationship between the latent variable and the estimated class probabilities, since the logistic functions fitted by this model on the latent variable axis respect the constraints given in Proposition 4.1. As a result, the fitted logistic curves are particularly useful to model data with an underlying model of type (12). Based on these findings, we propose an extension of the proportional odds model for learning partial class memberships in an ordinal setting. To this end, we define the logit of the cumulative partial membership vector as follows:

$$\text{logit}(\overline{f}_k(\mathbf{x})) = \log\Big(\frac{\overline{f}_k(\mathbf{x})}{1 - \overline{f}_k(\mathbf{x})}\Big)\,. \tag{14}$$

This logit can be modeled as a (linear) function of the features

$$\text{logit}(\overline{f}_k(\mathbf{x})) = \mathbf{w} \cdot \mathbf{x} + \theta_k\,, \tag{15}$$

for $k = 1, \ldots, K-1$.

Traditionally, proportional odds models are fit to the data through a likelihood maximization. The use of maximum likelihood implies probabilistic interpretation of the partial membership vectors, an assumption which is often no valid. However, the likelihood framework results in an optimization problem which can easily be solved to optimality by a gradient based algorithm. Moreover, experiments indicate that the loss function which is obtained within the likelihood framework can be considered as a reasonable approximation to L_{MAEC}.

4.4. Kernelized proportional odds

In its most basic form, the proportional odds model models the logits as linear functions of the input. In

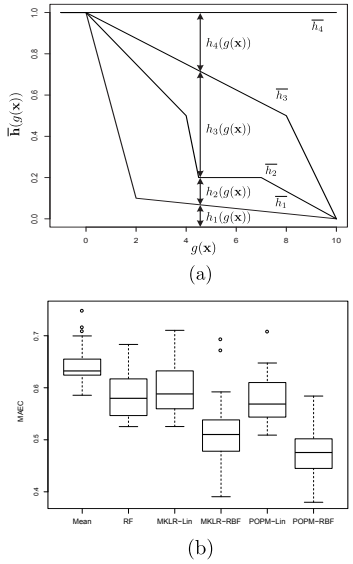

Figure 1. (a) Representation of the functions $\overline{h}_1, \ldots, \overline{h}_4$ in the experimental setup. (b) An overview of the performance in terms of L_{MAEC}.

case of partial memberships, we give two reasons why this is too restrictive.

First of all, there is no reason to assume that the latent variable $g(\mathbf{x})$ can be modeled through a linear function of the inputs. Often, the assumption of linearity is wrong. To overcome this problem, a kernelized version of the proportional odds model is proposed. Here, the logit of the k-th class is modeled as follows:

$$\text{logit}(\overline{f}_k(\mathbf{x})) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x}) + \theta_k \,,$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ a vector of parameters and $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ a Mercer kernel. Model parameters can be estimated through maximizing the L2-penalized maximum likelihood function.

Secondly, when the proportional odds model is used to predict partial memberships, $\overline{h}_1, \ldots, \overline{h}_{K-1}$ are approximated by logistic functions. In practice, there is no reason to assume that this will result in a good approximation. Experiments indicated that, for $\overline{h}_1, \ldots, \overline{h}_{K-1}$ strongly deviating from logistic functions, the increase in flexibility obtained through the kernelization resulted in an improved performance, even in case of a linear function for $g(\mathbf{x})$.

4.5. Experiments

In this section, the predictive capabilities of the kernelized proportional odds model for partial memberships (POPM) are demonstrated through an experiment on artificial data. In this experiment, we consider the setting of inferring partial ordinal class memberships with 7-dimensional inputs and 4-dimensional outputs, representing partial class memberships for 4 classes. The performance of POPM is compared to several methods. As a baseline method, the performance of the arithmetic mean as a predictor was considered. Other methods were the kernelized multi-class logistic regression model for partial memberships (MKLR) and a random forest regression model which was extended for partial ordinal class memberships (RF). For the kernelized methods, we tested the RBF kernel as well as the linear kernel.

The artificial data was generated as follows. Firstly, feature vectors were drawn from a uniform distribution over $[1, 2]^7$. Secondly, the partial membership vectors were generated using (12) following Proposition 4.2 where $s = 100$ with $\overline{h}_1, \ldots, \overline{h}_{K-1}$ as shown in Figure 1 and

$$g(\mathbf{x}) = 9\sqrt{\left|\frac{x_1 + x_2 - x_3 - x_4}{x_5 + x_6 - x_7}\right|} - 1.2\,.$$

To train the models, a training set containing 20 instances and a validation set containing 20 instances were created. To test their performance, a test set containing 1000 instances was created as well. The performance of each method in terms of the L_{MAEC} is shown in Figure 1. The boxplots in this figure are the result of 30 repetitions of the data generation process. It can be seen that POPM outperforms the other methods in this example.

5. Summary

We have extended the ordinal regression setting to enable it to deal with partial memberships. The principle of stochastic dominance was used to define a partial order on partial membership vectors. Finally, a kernelized version of the proportional odds model was proposed to increase the flexibility and applicability in practical situations.

References

Aitchison, J. (1986). *The statistical analysis of compositional data.* Chapman &Hall.

Anonymous, A. (2009). Suppressed for anonymity.

Levy, H. (2006). *Stochastic dominance, investment decision making under uncertainty, 2nd edition.* Springer.

McCullagh, P. (1980). Regression models for ordinal data. *J. R. Statist. Soc.*, *4*, 109–142.